FACULDADE DE INFORMÁTICA E ADMINISTRAÇÃO PAULISTA
CURSO DE PÓS-GRADUAÇÃO EM DATA ANALYTICS

JORGE DIEGO GUARDA GIBIN

Fase 3 – Big Data

CAMPINAS
2025

JORGE DIEGO GUARDA GIBIN

Fase 3 – Big Data

Trabalho apresentado a Faculdade de Informática e Administração Paulista, como entrega do Tech Challenge da Fase 3 – Big Data

CAMPINAS

2025

RESUMO

Este trabalho tem como objetivo analisar dados relacionados à pandemia de COVID-19 utilizando técnicas de Data Analytics, com foco em padrões de comportamento, características populacionais e sintomas da doença. A análise foi baseada na base de dados PNAD-COVID-19 do IBGE, que contém informações sobre sintomas, comorbidades e características socioeconômicas de uma amostra representativa da população brasileira. O estudo envolveu a aplicação do algoritmo K-means para a formação de clusters, visando identificar grupos com maior risco de complicações graves, além da Análise de Componentes Múltiplos (MCA) para reduzir a dimensionalidade dos dados. A pesquisa também utilizou métodos estatísticos para investigar correlações entre sintomas, comorbidades e internações. Os resultados destacam a importância de monitorar pacientes com comorbidades e aqueles com mais de 30 anos, especialmente os acima de 60 anos, recomendando ações como monitoramento remoto, vacinação prioritária e direcionamento eficaz de recursos. Este trabalho oferece uma base sólida para otimizar a gestão hospitalar e preparar-se para futuros surtos, contribuindo para um melhor planejamento e resposta às crises de saúde pública.

SUMÁRIO

1 Enunciado

Imagine agora que você foi contratado(a) como Expert em Data Analytics por um grande hospital para entender como foi o comportamento da população na época da pandemia da COVID-19 e quais indicadores seriam importantes para o planejamento, caso haja um novo surto da doença.

Apesar de ser contratado(a) agora, a sua área observou que a utilização do estudo do PNAD-COVID 19 do IBGE seria uma ótima base para termos boas respostas ao problema proposto, pois são dados confiáveis. Porém, não será necessário utilizar todas as perguntas realizadas na pesquisa para enxergar todas as oportunidades ali postas.

É sempre bom ressaltar que há dados triviais que precisam estar no projeto, pois auxiliam muito na análise dos dados:

• Características clínicas dos sintomas;

• Características da população;

• Características econômicas da sociedade.

O Head de Dados pediu para que você entrasse na base de dados do PNAD-COVID-19 do IBGE (https://covid19.ibge.gov.br/pnad-covid/) e organizasse esta base para análise, utilizando Banco de Dados em Nuvem e trazendo as seguintes características:

- a. Utilização de no máximo 20 questionamentos realizados na pesquisa;
- b. Utilizar 3 meses para construção da solução;
- c. Caracterização dos sintomas clínicos da população;
- d. Comportamento da população na época da COVID-19;
- e. Características econômicas da Sociedade;
- Seu objetivo será trazer uma breve análise dessas informações, como foi a organização do banco, as perguntas selecionadas para a resposta do problema e

quais seriam as principais ações que o hospital deverá tomar em caso de um novo surto de COVID-19.

2 Projeto

Para a criação desta análise, o projeto foi dividido em 4 etapas:

- Análise dos dados disponíveis
- Análise exploratória de dados
- Análise estatística
- Conclusão

3 Análise dos dados disponíveis

Para acessar a base de dados da PNAD-COVID do IBGE de forma eficiente, considerando seu grande volume de informações, optou-se por utilizar um banco de dados em nuvem, pois ao ler a documentação, foi identificado que as bases de dados necessárias para análises estão disponíveis em um conjunto de dados públicos dentro da ferramenta de dados BigQuery. Optar por manter os dados dentro da plataforma de dados garante escalabilidade, rapidez nas consultas e acesso remoto, otimizando o processo de análise e garantindo caso haja uma atualização no banco, este dado seja replicado de forma automática para as consultas.

As bases de dados contêm informações obtidas por meio de um questionário aplicado a uma amostra da população. Para a análise foram selecionadas 18 perguntas do questionário original no intervalo de tempo com início em 09/2020 e final em 11/2020, totalizando 3 meses, período este que apresenta uma queda no percentual da população que declarou ter tido algum sintoma relacionado à síndrome gripal e que antecedeu o pico de mais mortes por COVID-19 no Brasil.

As 18 perguntas selecionadas foram:

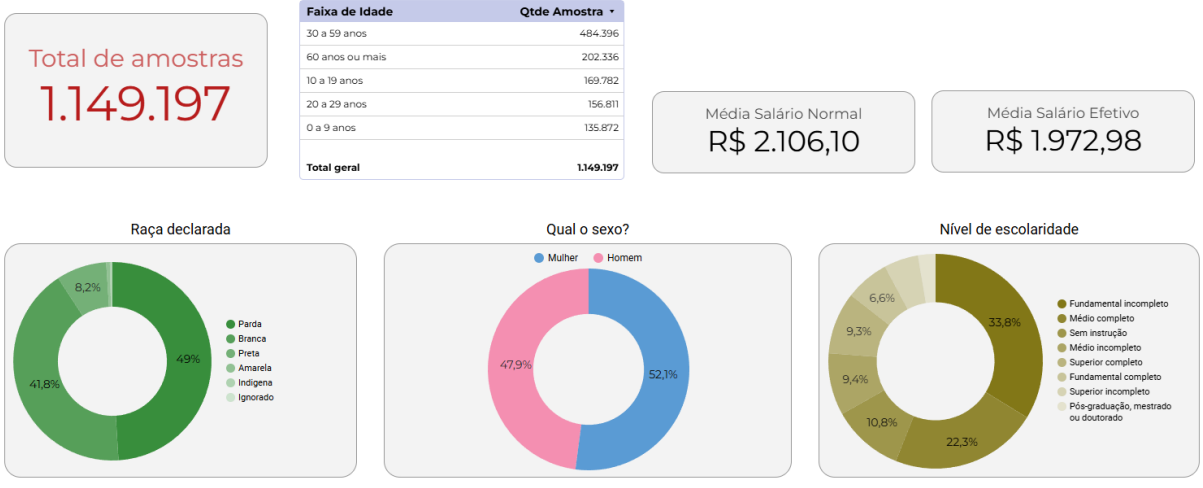
1. Qual sua idade?
2. Qual sexo?
3. Cor ou raça
4. Qual sua escolaridade?
5. Na semana passada, o(a) sr(a) teve quais sintomas?
6. Por causa disso, o(a) sr(a) foi a algum estabelecimento de saúde (na semana passada)?
7. Que providência o(a) sr(a) tomou para se recuperar desses sintomas (na semana passada)?

8. Em que local buscou o atendimento (na semana passada)?
9. Ao procurar o estabelecimento de saúde, teve que ficar internado(a) por um dia (24 horas) ou mais?
10. Durante esta internação, o(a) sr(a) foi sedado(a), entubado(a) e colocado(a) em respiração artificial com ventilador?
11. O(A) sr(a) tem algum plano de saúde médico, seja particular, de empresa ou de órgão público?
12. O(A) sr(a) fez algum teste para saber se estava infectado(a) pelo coronavírus?
13. Qual tipo de teste utilizado (swab nasofaríngeo, sangue dedo ou veia).
14. Algum médico já lhe deu o diagnóstico de alguma dessas doenças?
15. Na semana passada, devido à pandemia do Coronavírus, em que medida o(a) sr(a) restringiu o contato com as pessoas?
16. Na semana passada, por pelo menos uma hora, o(a) sr(a) trabalhou ou fez algum bico?
17. Quanto recebeu (ou retira) normalmente em seu trabalho?
18. Quanto recebeu (ou retirou) efetivamente em seu trabalho?

Foram utilizados scripts em SQL que estão salvos dentro do projeto para a realização de todo ETL, para a análise exploratória foi criado um dashboard em Looker Studio para facilitar a visualização dos dados e para análise estatística foi utilizado notebook Python dentro do BigQuery a fim de centralizar os dados em um único local.

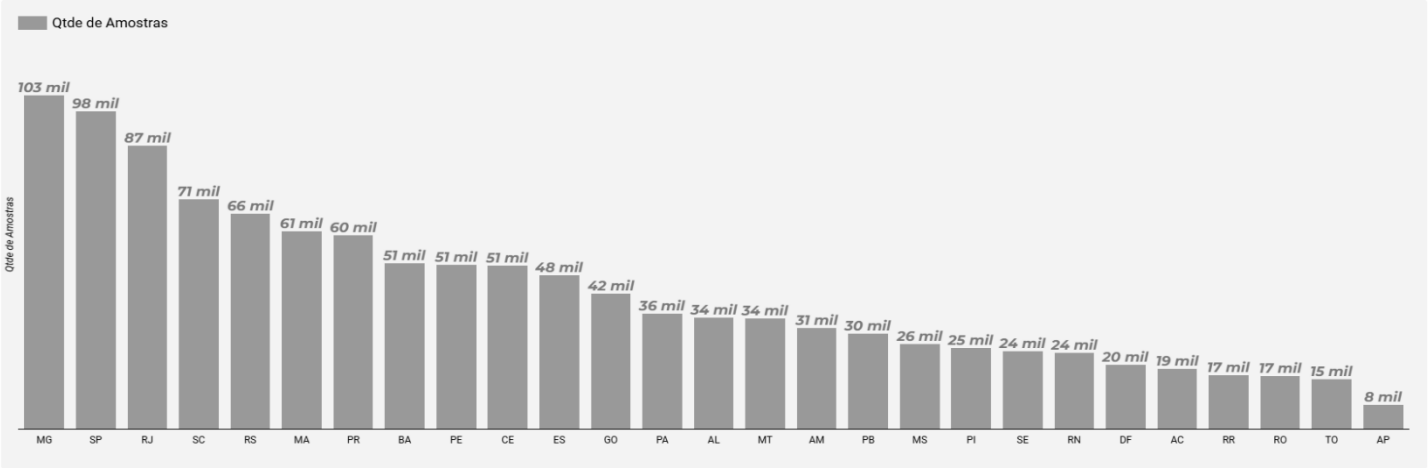
4 Análise descritiva dos dados

Para a análise descritiva inicialmente foi mapeado quais eram as características da população entrevistada no período, utilizando as perguntas: qual o sexo, raça e cor, escolaridade, faixa etária, quanto recebeu (ou retira) normalmente em seu trabalho e quanto recebeu (ou retirou) efetivamente em seu trabalho, obtendo os seguintes dados:



Com esta análise foi possível observar que no período houveram 1.149.197 de entrevistados, cerca de 0,55% da população brasileira em 2020, havia uma pariedade na quantidade de homens e mulheres entrevistados, mais 59% dos entrevistados estavam na faixa de idade com trinta ou mais anos, 45% da população não havia concluído o ensino fundamental e 49% da população se declara parda. Na análise socioeconômica, foi possível concluir que os entrevistados declararam um salário normal de 2 salários minímos (R$ 1.045,00) e o salário que efetivamente recebido ficou abaixo de 2 salários minímos.

Para obter mais informações, fizemos uma análise geográfica para entender em quais estados foram colhidas as entrevistas a fim de análisar se era possível clusterizar a análise por região do país, o que ficou evidenciado na análise macro que não era possível seguir desta forma, uma vez que há distribuição por região estava dispersa.

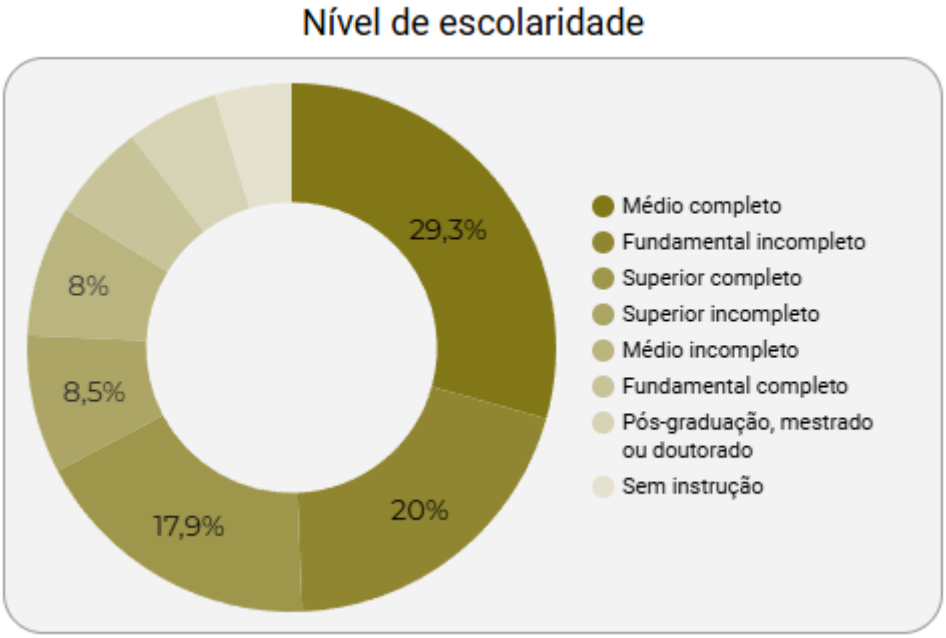


Foi realizada a análise também se o entrevistado teve sintomas, tento o resultado surpreendido, uma vez que mais de 95% dos entrevistados relataram não ter tido sintomas.

Foi realizada a análise também pela quantidade de amostras que testaram positivo para COVID-19 e isto nos revelou uma mudança em dois cenários, faixa etária conforme imagem abaixo, onde há um aumento na faixa entre 20 e 29 anos.

Faixa de Idade	Qtde Amostra
30 a 59 anos	19.356
20 a 29 anos	5.238
60 anos ou mais	4.853
10 a 19 anos	2.338
0 a 9 anos	1.137
Total geral	**32.922**

E há uma mudança também no nível de escolaridade, onde a maioria dos entrevistados que testaram positivo para COVID-19 estão entre as classes de Ensino Médio Completo e Ensino Superior Completo e Incompleto.

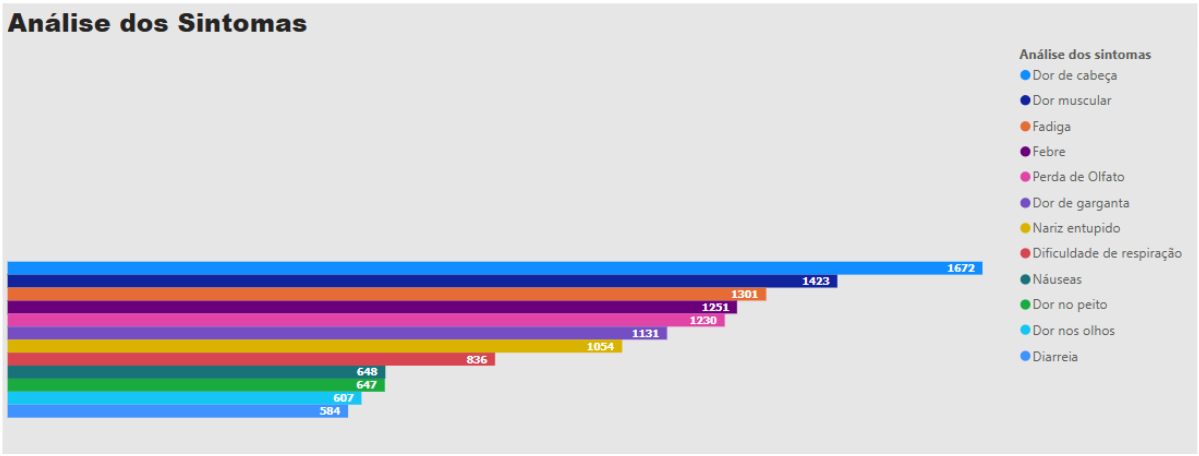


Após as análises e visando celeridade na resolução e abordando os grupos etários com mais casos e classificado como de risco, clusterizamos as análises as faixas etárias de 30 anos ou mais, com foco somente nos casos que foram testados positivos para COVID-19, uma vez que este estudo visa explicitar padrões para que a instituição de saúde possa atender de forma eficaz a população brasileira. Esta decisão é corroborada utilizando uma matéria publicada (https://portal.fiocruz.br/noticia/estudo-analisa-registro-de-obitos-por-covid-19-em-2020#:~:text=Em%202020%2C%20ocorre-

ram%201.207%20óbitos,de%2028%20dias%20de%20vida) no site da Fundação Oswaldo Cruz, renomada instituição criada em 25 de maio de 1900 - com o nome de Instituto Soroterápico Federal -, onde possui a missão de combater os grandes problemas da saúde pública brasileira. Para isso, moldou-se ao longo de sua história como centro de conhecimento da realidade do País e de valorização da medicina experimental e teve participação crucial no combate a COVID-19 no Brasil e no mundo.

4.1 Análise dos casos positivos de COVID-19 para faixa de 30 anos ou mais

Iniciando a análise, considerando somente casos positivos para entrevistados onde a faixa etária é de 30 anos ou mais, fizemos uma análise para saber quais eram os principais sintomas declarados pelos entrevistados, para assim entender como a doença se comporta:



Dor de cabeça, foi o sintoma mais frequente com **12%,** porém ao analisarmos por faixa etária, percebemos uma diferença nos sintomas apresentados onde pessoas mais de 60 anos os sintomas de **dor muscular, fadiga e tosse** foram os sintomas mais frequentes.

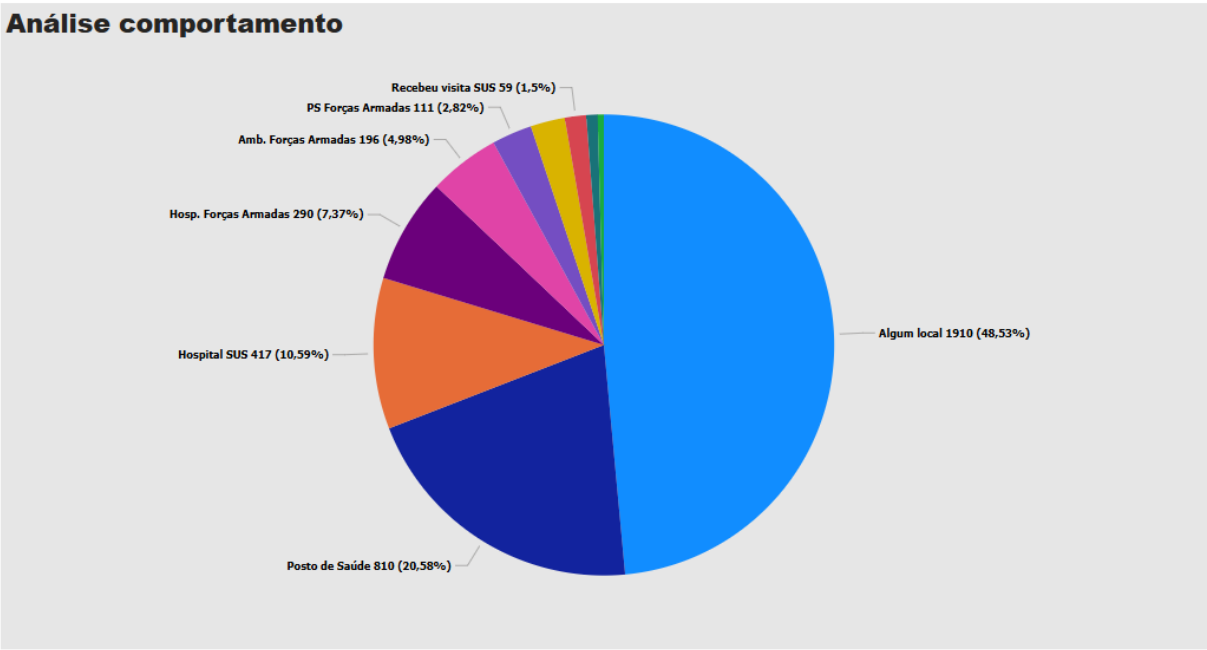
Ainda na análise de sintomas, foi feita uma análise se haviam entrevistados com comorbidade e como elas se comportavam no cenário, sendo que 67% são hipertensos

e/ou diabéticos, o que revela que é necessário ter uma atenção especial a estes grupos crônicos, pois são os mais afetados.

Com as análises acima concluímos que os grupos com mais de 30 anos e com doenças crônicas como hipertensão e diabetes são os mais suscetíveis a contrair o SARS-CoV-2, e para estes o monitoramento remoto, vacinação prioritária e educação e orientação são algumas ações que as unidades de atendimento médico podem ter para prevenir que estes casos evoluam a óbito.

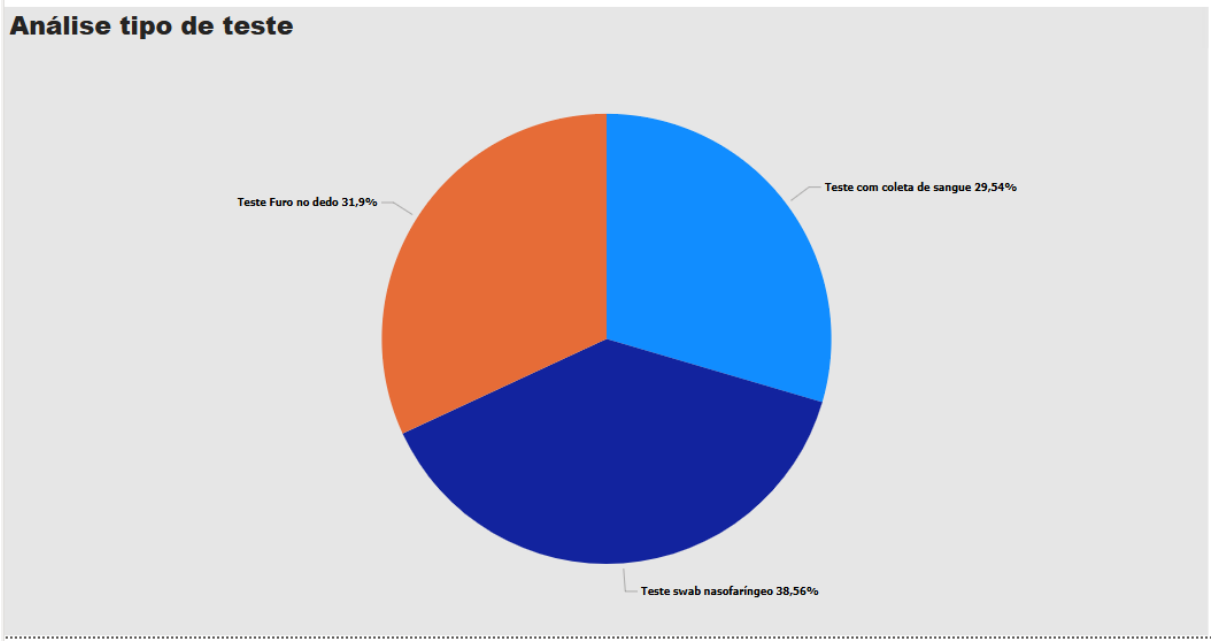
4.2 Análise de ações tomadas pelos entrevistados

A análise abaixo visa identificar quais foram as ações tomadas pelos entrevistados a fim de verificar um padrão e tendências, para que em casos futuros as unidades de saúde possam se antecipar.



Podemos verificar que entre as pessoas que **não** buscaram atendimento, **ficar em casa**, foi a principal providência tomada para cuidar dos sintomas, conforme mostramos nos gráficos acima, o que reflete a orientação das autoridades médicas na época dos fatos analisados.

Fizemos também a análise de quais métodos foram utilizados para coleta de exame:

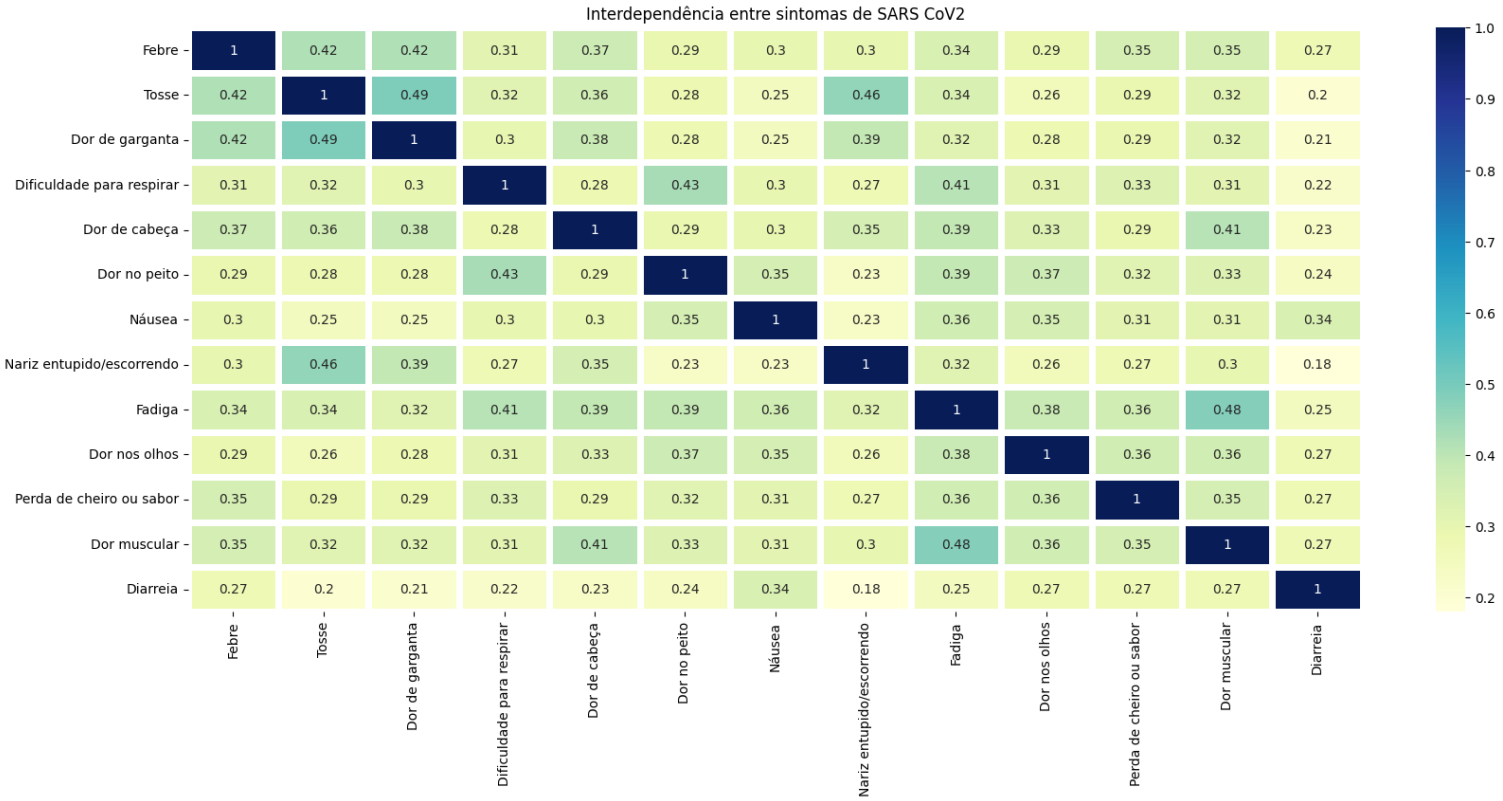


Vemos no slide acima que não há uma preferência pelo tipo de coleta, sendo usado de forma praticamente igual os três métodos.

Analisamos também que somente 24% dos pacientes que foram internados precisaram ser sedados.

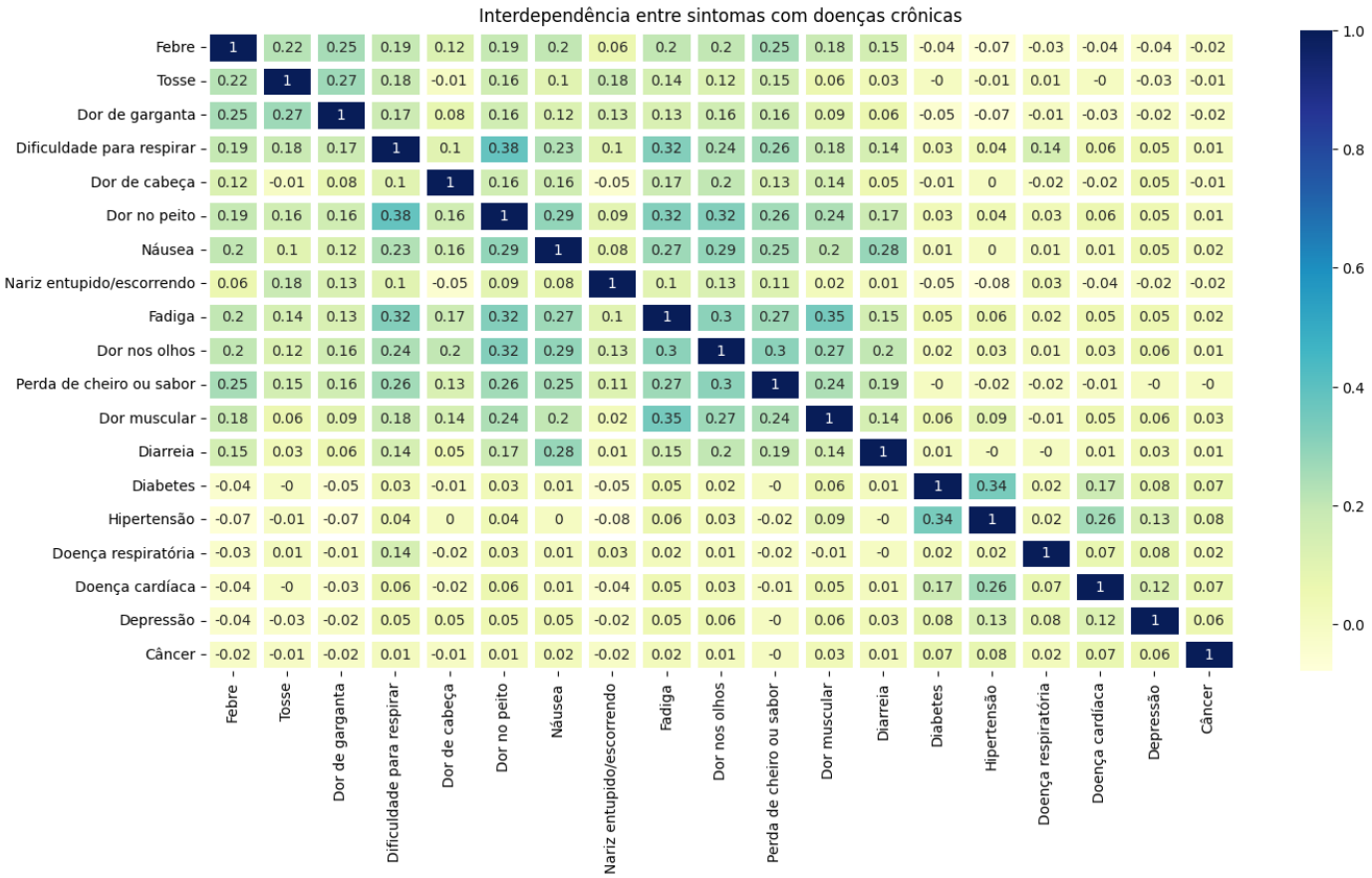
5 Análise estatística

A análise de correlação desempenha um papel essencial na compreensão das relações entre variáveis, utilizando métodos estatísticos para identificar como elas se interconectam. Foi realizado três tipos principais de análise nesse contexto: a primeira avaliou a interdependência entre os diversos sintomas da COVID-19; a segunda examinou a relação entre esses sintomas e a presença de doenças crônicas pré-existentes e, por fim, a terceira investigou a associação entre os sintomas e os casos de internação e intubação. Para facilitar as correlações, foi necessário converter as variáveis categóricas em valores booleanos, considerando como true as respostas positivas para os sintomas e false para as demais opções.

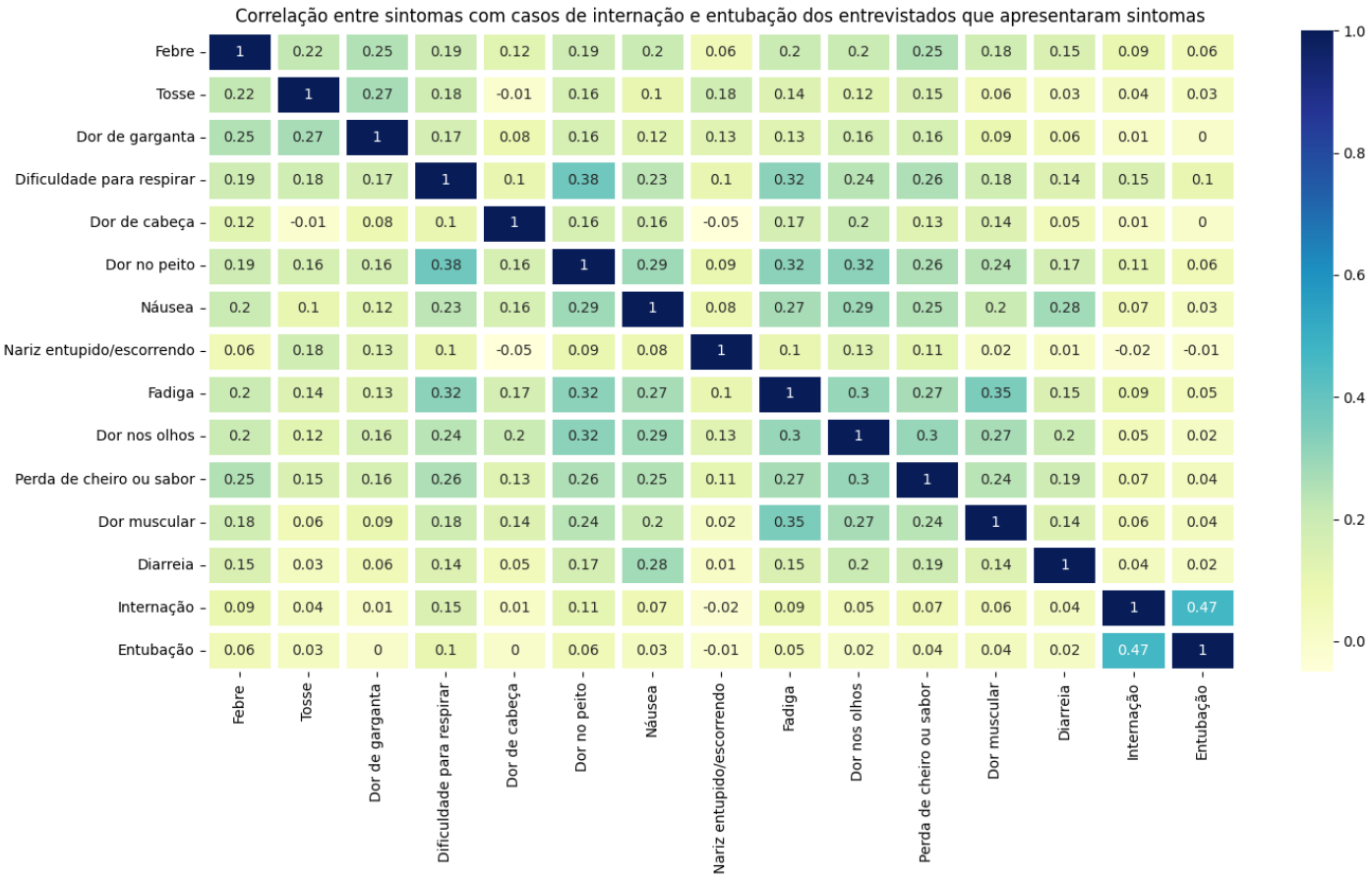


O heatmap revela as correlações entre diversos sintomas da COVID-19, destacando associações moderadas que são clinicamente relevantes. Entre as mais notáveis, encontramos a conexão entre dor de garganta e tosse, dor muscular e fadiga, dor de cabeça e dor muscular, febre e tosse, febre e dor de garganta, tosse e nariz entupido/escorrendo, além da associação entre dificuldade para respirar e dor no peito, e dificuldade para respirar e fadiga.

Essas correlações seguem uma lógica clínica, onde o agravamento de um sintoma frequentemente está associado ao surgimento ou intensificação de outro, refletindo a progressão típica da doença. Esse conhecimento é crucial para o planejamento e gestão de cuidados, pois permite antecipar combinações de sintomas em pacientes contaminados com SARS-CoV2. Com base nesses insights, é possível otimizar a preparação clínica e logística, garantindo uma resposta mais eficiente e direcionada às necessidades dos pacientes.



A segunda análise de correlação investiga a relação entre os sintomas da COVID-19 e a presença de doenças crônicas, o que é crucial para compreender como essas condições podem influenciar a manifestação clínica dos sintomas. Embora as doenças crônicas não estejam diretamente associadas ao surgimento de novos sintomas, os dados indicam que elas podem exacerbar sintomas preexistentes, resultando em um quadro clínico de maior gravidade. Esse fator representou uma preocupação substancial durante a pandemia, evidenciando a necessidade de monitoramento rigoroso de pacientes com comorbidades, que apresentaram um risco significativamente elevado de complicações graves.



A análise de correlação entre sintomas e casos de internação e intubação não indicou que um sintoma isolado seja diretamente responsável por tais intervenções clínicas. Contudo, é fundamental considerar os contextos clínicos observados nos hospitais durante a pandemia, os quais revelaram uma preocupação crescente com pacientes que apresentavam doenças crônicas, comorbidades e idade avançada.

Embora esses fatores não estejam necessariamente vinculados a sintomas específicos, eles amplificam o risco de internação e intubação devido a fragilidades biológicas preexistentes. Tais dados ressaltam a vulnerabilidade dessas populações, que demonstraram uma maior propensão a desenvolver complicações graves associadas à COVID-19, evidenciando a necessidade de uma abordagem mais cuidadosa e estratégica no manejo de pacientes com essas características

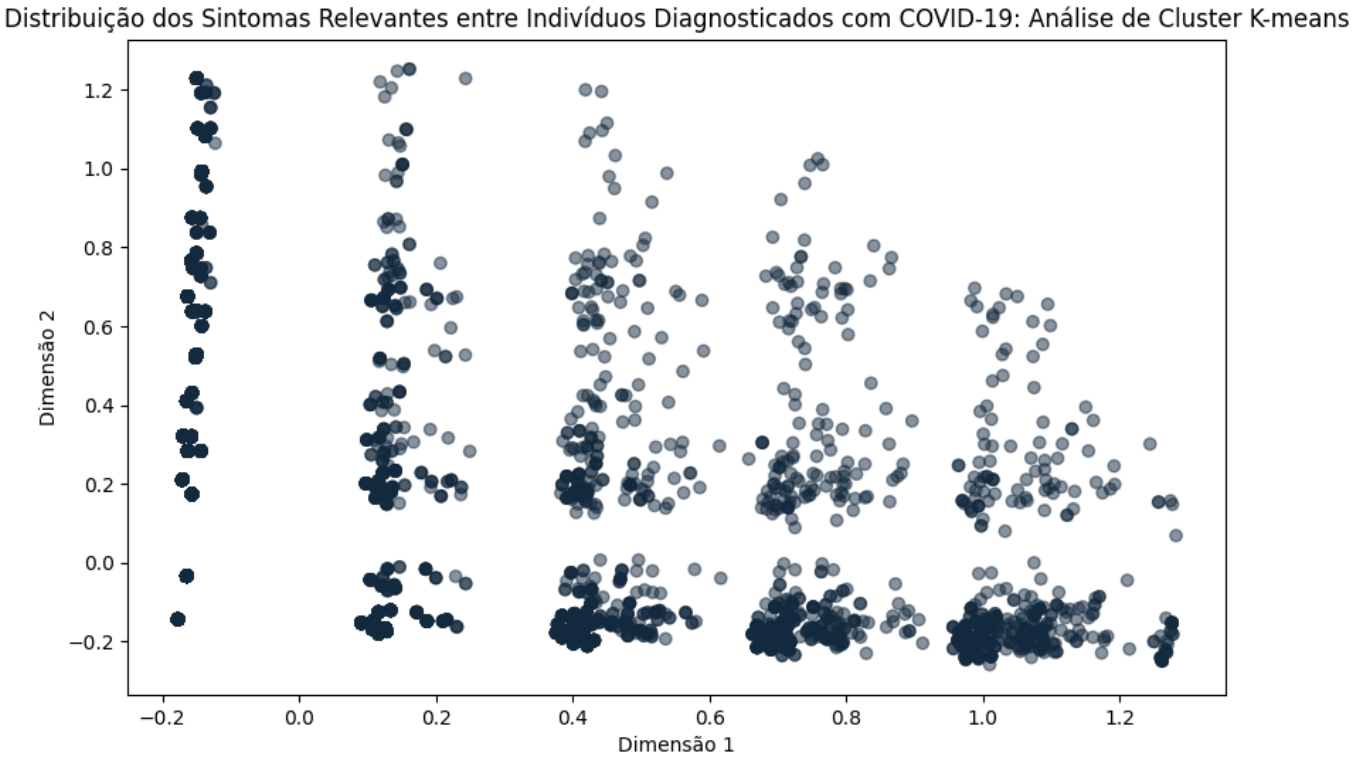
6 Análise de Agrupamento utilizando o Algoritmo K-means

A aplicação do algoritmo de cluster K-means aos dados teve como objetivo identificar os sintomas mais relevantes em indivíduos diagnosticados com a COVID-19. Abaixo, apresenta-se uma descrição detalhada dos passos seguidos para realizar essa análise:

Seleção dos Casos Positivos: O primeiro passo no processo de análise foi a filtragem dos dados para isolar apenas os casos positivos para a COVID-19. Foram considerados positivos os indivíduos que apresentaram resultados positivos em pelo menos um dos três tipos de testes disponíveis na base de dados da PNAD: teste de swab nasal ou oral, exame de sangue por punção no dedo ou exame de sangue por punção venosa.

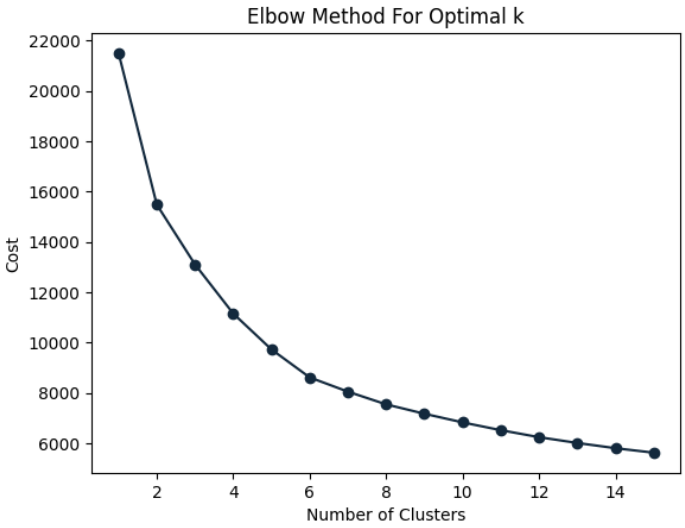
Tratamento dos Dados Categóricos: Os dados categóricos relativos aos sintomas foram convertidos em variáveis binárias, com a atribuição do valor "true" para a presença de um sintoma (indicando uma resposta positiva) e "false" para sua ausência (demais respostas). Esse tratamento foi realizado para padronizar os dados, permitindo que as variáveis fossem adequadamente utilizadas nas etapas seguintes da análise.

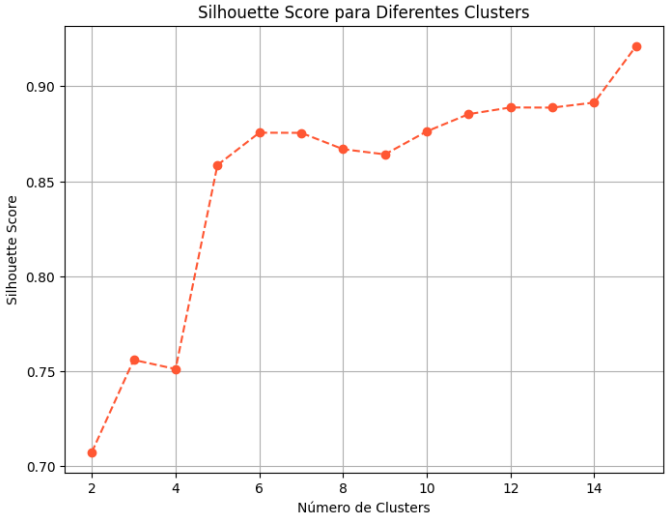
Análise de Componentes Múltiplos (MCA): Após a filtragem e o tratamento dos dados, foi aplicada a técnica de Análise de Componentes Múltiplos (MCA), com o objetivo de reduzir a dimensionalidade dos dados categóricos. A MCA permitiu identificar as variáveis que mais contribuem para a variabilidade entre os casos analisados. Essas variáveis foram então organizadas em duas dimensões principais, facilitando a interpretação e análise dos agrupamentos gerados pelo algoritmo K-means.



Remoção de Outliers: Durante a realização da Análise de Componentes Múltiplos (MCA), foram identificados outliers com base na distância euclidiana e removidos da análise. Esses pontos discrepantes podem interferir na formação dos clusters, comprometendo a obtenção de resultados consistentes e conclusões significativas. A equipe responsável pela análise foi alertada sobre a importância de monitorar esses outliers, uma vez que eles podem refletir características raras ou específicas, as quais poderiam justificar uma investigação mais aprofundada.

Determinação do Número Ideal de Clusters: Para identificar o número adequado de clusters, foram aplicados os métodos de cotovelo (Elbow Method) e Silhouette. Além dessas abordagens quantitativas, também foi realizada uma avaliação visual da dispersão dos pontos no espaço da análise MCA, a fim de corroborar os resultados obtidos pelos métodos formais e garantir uma segmentação mais precisa e representativa dos dados.





Aplicação do K-Means: Com base nos resultados obtidos na etapa anterior, determinamos que o número ideal de clusters para o conjunto de dados analisado era 15. A partir dessa definição, aplicou-se o algoritmo K-means aos dados previamente transformados por meio da Análise de Componentes Múltiplos (MCA). Este procedimento resultou na formação de agrupamentos que possibilitaram a observação de padrões significativos dentro de cada cluster, facilitando a identificação de tendências e características comuns entre os grupos. A análise desses clusters proporcionou uma compreensão mais detalhada e estruturada dos dados, essencial para a extração de insights relevantes.

cluster	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
counts	20525	261	1421	322	570	112	570	3567	323	221	460	181	1406	147	1189

Agrupamento de Clusters e Cálculos Estatísticos: Para cada cluster gerado, calculou-se a mediana da relevância das variáveis resultantes da Análise de Componentes Múltiplos (MCA), além do desvio padrão das duas dimensões obtidas na análise. Esse procedimento permitiu a identificação de clusters mais homogêneos e revelou aqueles com maior relevância para as variáveis estudadas, proporcionando uma visão mais clara sobre a distribuição dos dados e a significância de cada agrupamento.

x	median	-0.18	-0.15	-0.17	0.71	-0.16	0.51	0.12	-0.17	1.03	0.12	0.41	0.97	-0.16	0.43	-0.16
	std	0.00	0.07	0.01	0.05	0.06	0.17	0.03	0.01	0.09	0.03	0.04	0.17	0.06	0.08	0.00
y	median	-0.14	1.23	0.32	-0.17	0.76	0.70	-0.15	0.21	-0.19	0.23	-0.17	0.21	0.68	0.22	-0.03
	std	0.00	0.06	0.04	0.04	0.09	0.15	0.04	0.01	0.05	0.10	0.04	0.13	0.05	0.08	0.00

Mediana e Desvio Padrão das Dimensões MCA por Cluster

cluster		1.00	4.00	5.00	12.00	3.00	5.00	8.00	11.00
x	median	-0.15	-0.16	0.51	-0.16	0.71	0.51	1.03	0.97
	std	0.07	0.06	0.17	0.06	0.05	0.17	0.09	0.17
y	median	1.23	0.76	0.70	0.68	-0.17	0.70	-0.19	0.21
	std	0.06	0.09	0.15	0.05	0.04	0.15	0.05	0.13

Combinação das Dimensões X e Y na Análise de Componentes Múltiplos (MCA)

CONCLUSÃO

A análise dos dados provenientes da PNAD-COVID-19 do IBGE revelou insights valiosos sobre o comportamento da população durante a pandemia de COVID-19 no período analisado, destacando padrões de sintomas, comorbidades, e características socioeconômicas. Através de uma abordagem estruturada que envolveu a análise descritiva, estatística e a aplicação de técnicas avançadas como o K-means e a Análise de Componentes Múltiplos (MCA), foi possível identificar subgrupos dentro da população que apresentam maior risco de complicações graves da doença.

Os principais achados indicam que indivíduos com comorbidades, como hipertensão e diabetes, e aqueles com mais de 30 anos, especialmente acima de 60, estão mais suscetíveis a manifestações graves da COVID-19, o que exige uma abordagem clínica diferenciada. A análise de correlação entre sintomas e internamento também evidenciou a relevância de monitorar atentamente esses pacientes, especialmente os que apresentam múltiplos sintomas associados à progressão da doença.

Além disso, a aplicação do algoritmo K-means permitiu a segmentação dos casos em clusters com base em características clínicas e comportamentais, facilitando a identificação de padrões significativos para o planejamento hospitalar. A combinação das dimensões MCA forneceu uma visão abrangente sobre os fatores que mais influenciam a evolução dos casos, permitindo um direcionamento mais eficiente dos recursos médicos.

Com base nos resultados obtidos, recomenda-se que os hospitais adotem estratégias focadas na monitorização remota, na vacinação prioritária e na educação da população de risco. A análise desses dados fornece uma base sólida para otimizar a gestão de recursos, antecipando necessidades futuras e ajudando a instituição de saúde a se preparar adequadamente para possíveis surtos de COVID-19 ou outras doenças similares. Assim, este estudo não apenas contribui para o enfrentamento imediato da pandemia, mas também oferece um framework para a gestão de crises de saúde pública a longo prazo.